Connecting the Canvas

Inside the Project LUX Linked Art Data Pipeline.

NotebookLM

The researcher's nightmare is a scattered puzzle

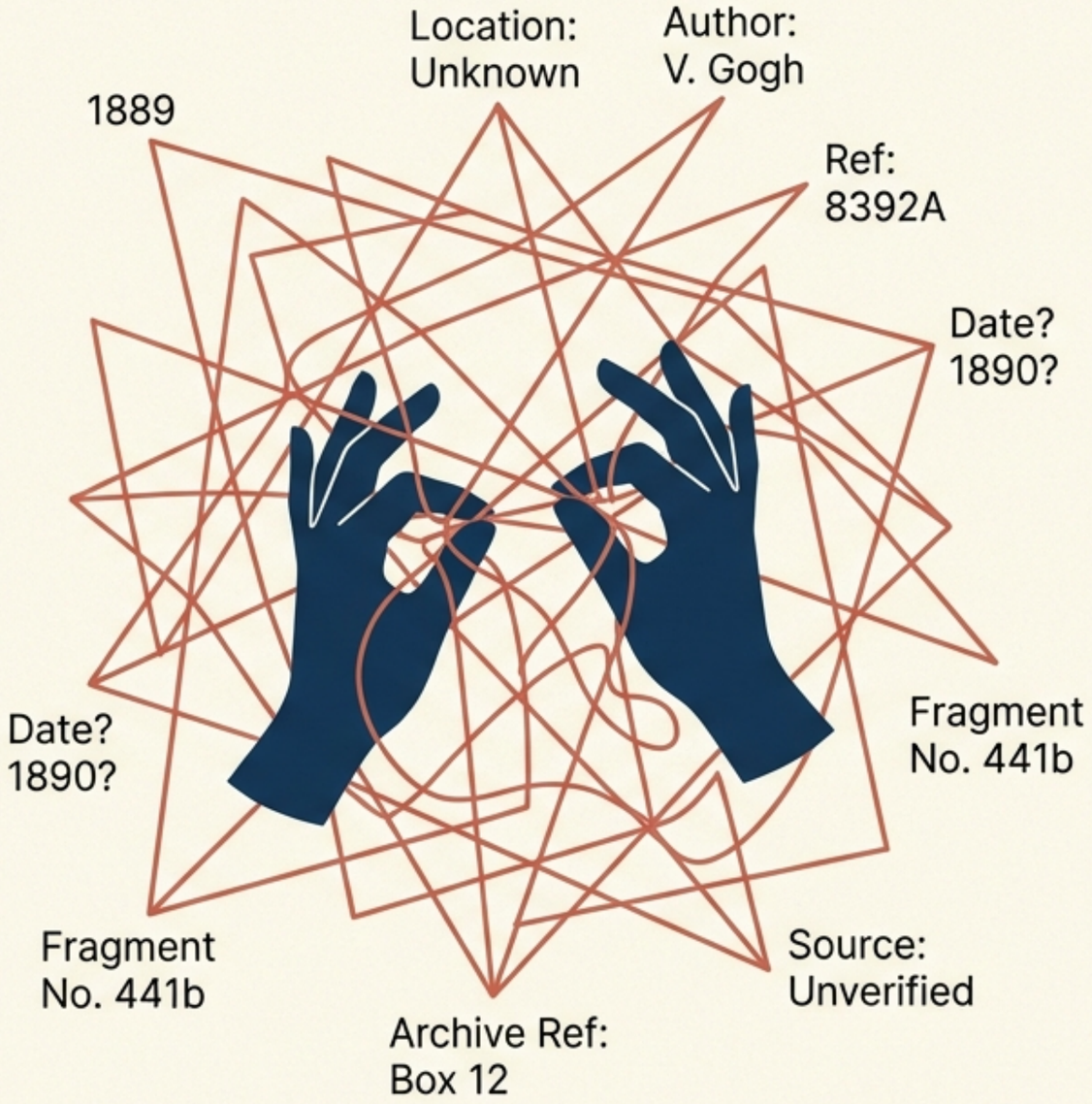
To get a complete picture of an artist, researchers must navigate paintings in one museum database, letters in a separate archive, and historical details in entirely different formats.

Crucial information exists in strict isolation, speaking different languages.

NotebookLM

Siloed data mandates endless manual wrangling

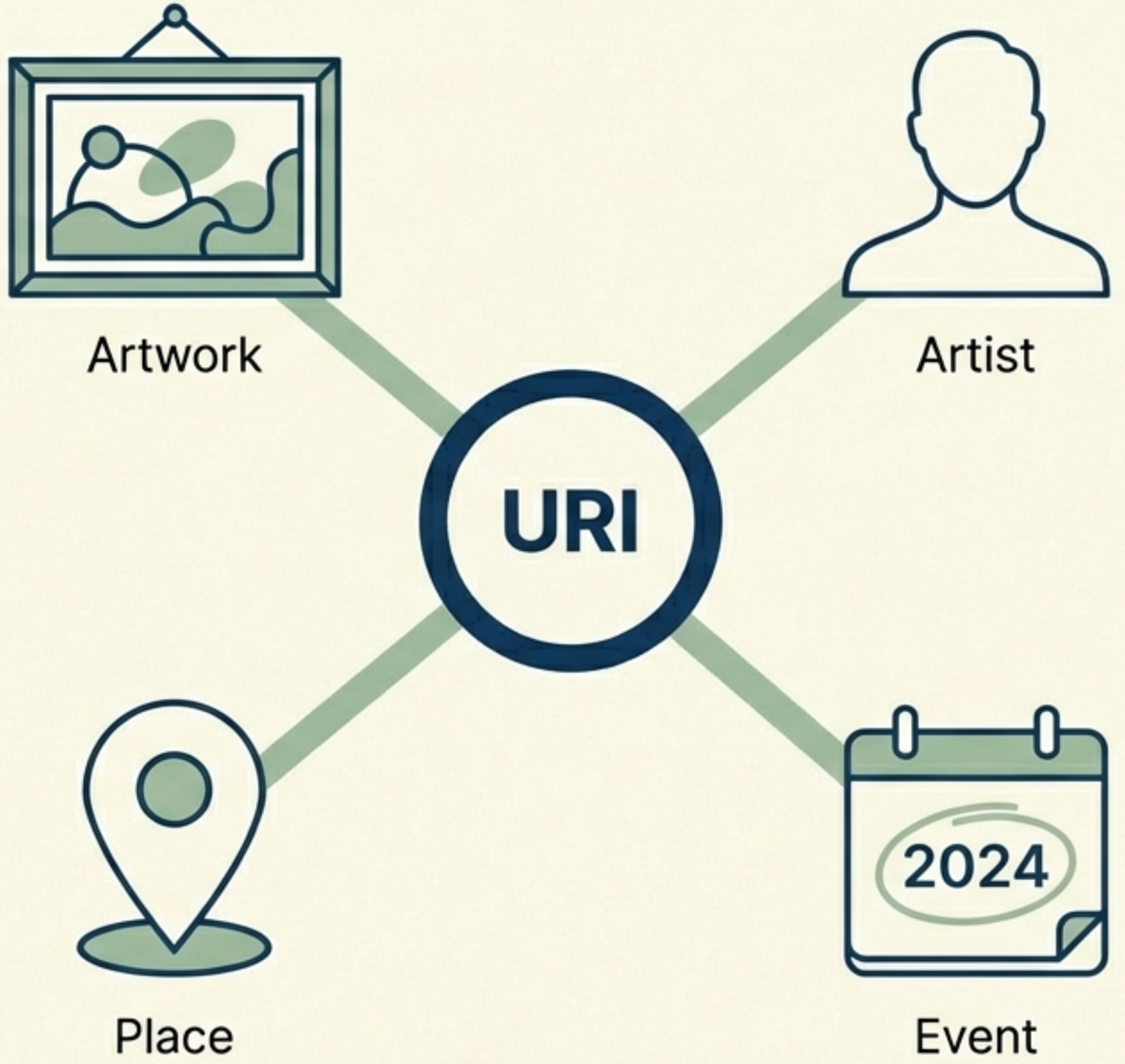
Without a unified system, researchers spend months untangling conflicting info and manually piecing together narratives. This approach leads to missed connections, redundant labor, and a focus on data cleanup rather than actual historical analysis.



NotebookLM

Linked Art provides a universal language

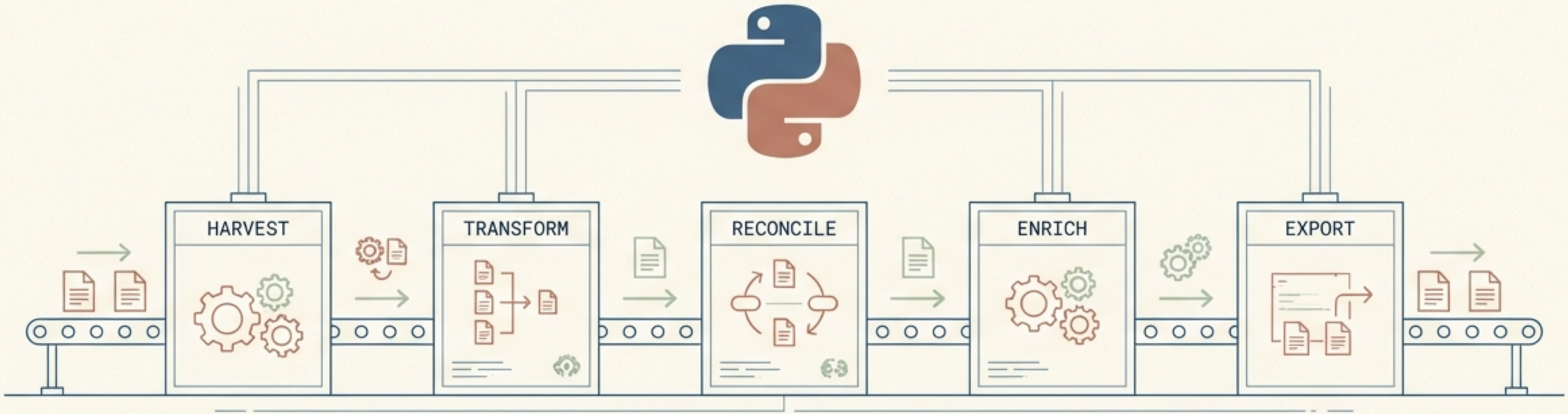
Linked Art creates a standard model for structuring cultural heritage information. By giving every piece of data its own permanent web address, machines can understand the relationships, allowing for sophisticated searches and new discoveries.



NotebookLM

The Project LUX automated data factory

Built entirely in Python (98.8%) with specialized shell scripting, the Project LUX data pipeline is a modular powerhouse. Its mission is to systematically harvest, transform, reconcile, enrich, and export Linked Art data into one cohesive, deeply interconnected knowledge base.

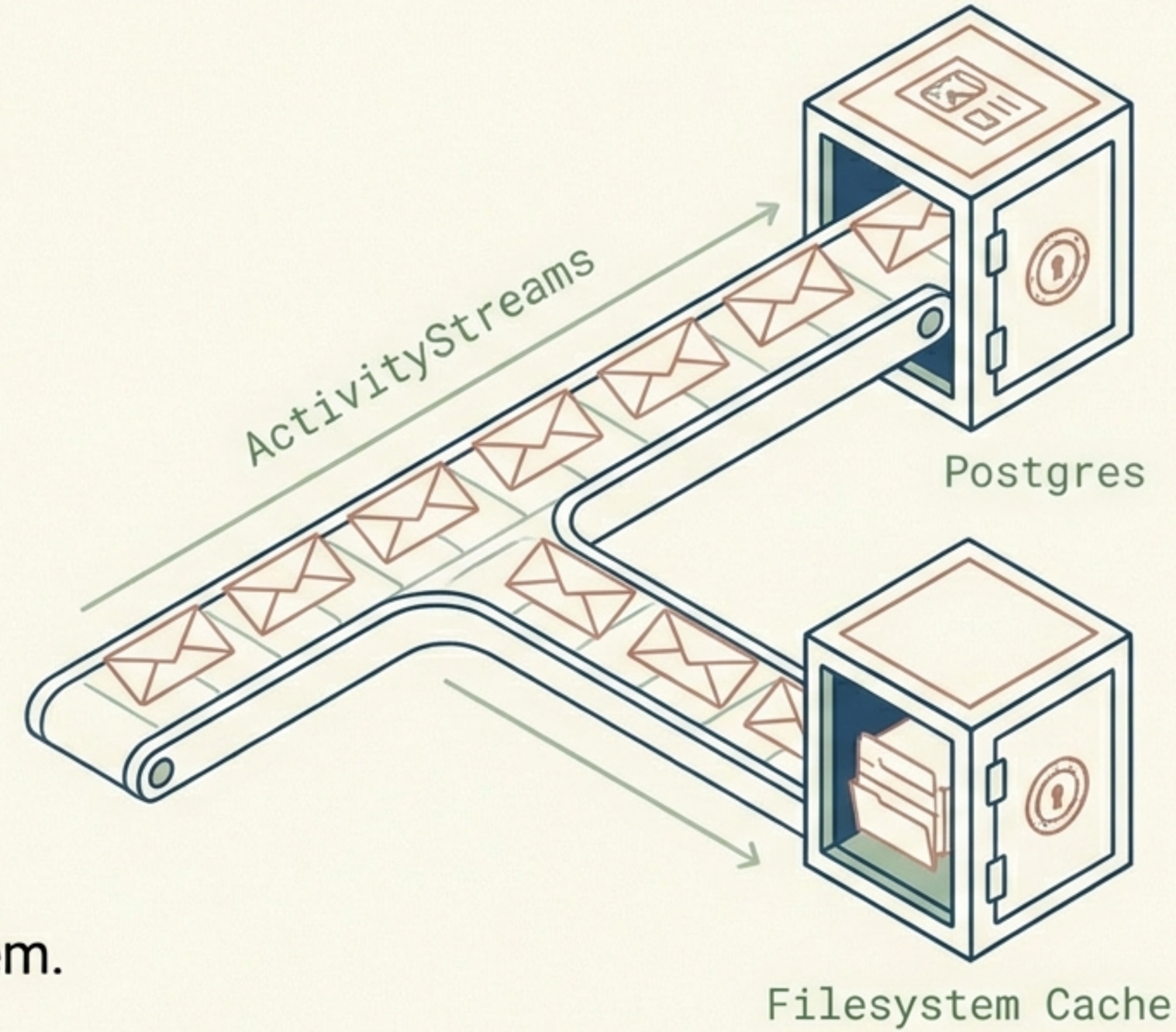


NotebookLM

Receiving and storage acts as a digital librarian

The Harvester component continuously scans for new or updated records.

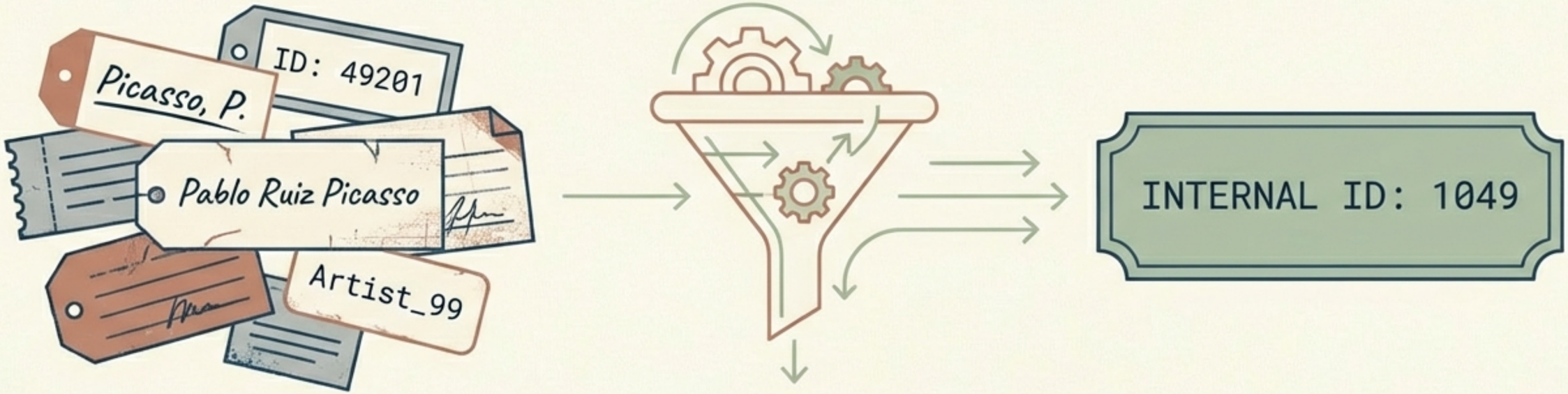
updated records. It relies on ActivityStreams—a standardized newsfeed for art data changes—to know exactly when relevant information is added. This raw data temporarily lands in a reliable Cache, utilizing Postgres or the local filesystem.



NotebookLM

The IdMap mints a consistent internal identity

The Identifier Map (IdMap) serves as the universal translator for names and numbers. It ensures that an entity recognized by one museum is correctly linked to the same entity from another source. It functions as a unique social security number within the system, regardless of what original identifier was used.



NotebookLM

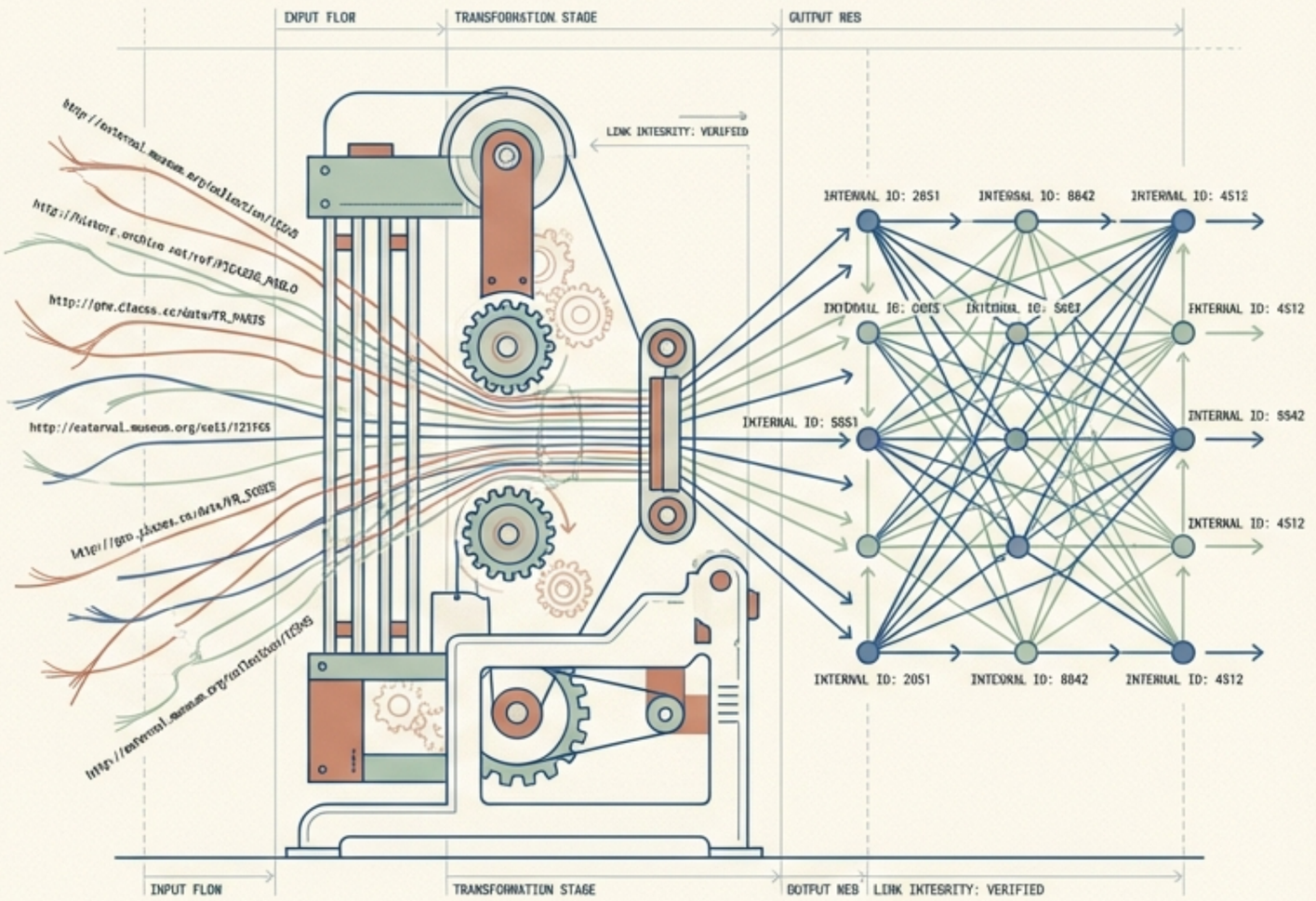
The core processing engine eliminates redundancy

- **Collector:** Recursively gathers every related identifier for a piece of data to ensure nothing is missed.
- **Reconciler:** Acts as a meticulous editor, checking for consistency across the facts.
- **Merger:** The unsung hero. It intelligently combines slightly different, disparate records representing the same entity into one single, clean, definitive entry.

NotebookLM

Weaving the web of internal connections

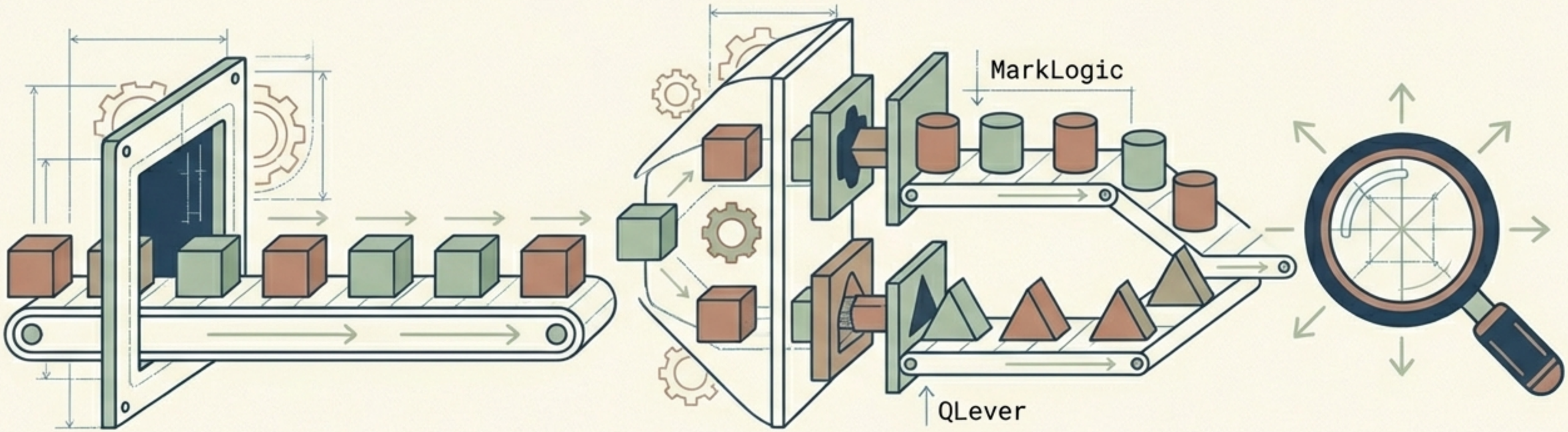
The Reidentifier recursively rewrites any external URIs within a record to use the consistent internal identifiers minted by the IdMap. Simultaneously, the ReferenceManager oversees the web of links, ensuring all connections between historical figures, places, and art are properly established and maintained.



NotebookLM

The Mapper translates output for discovery

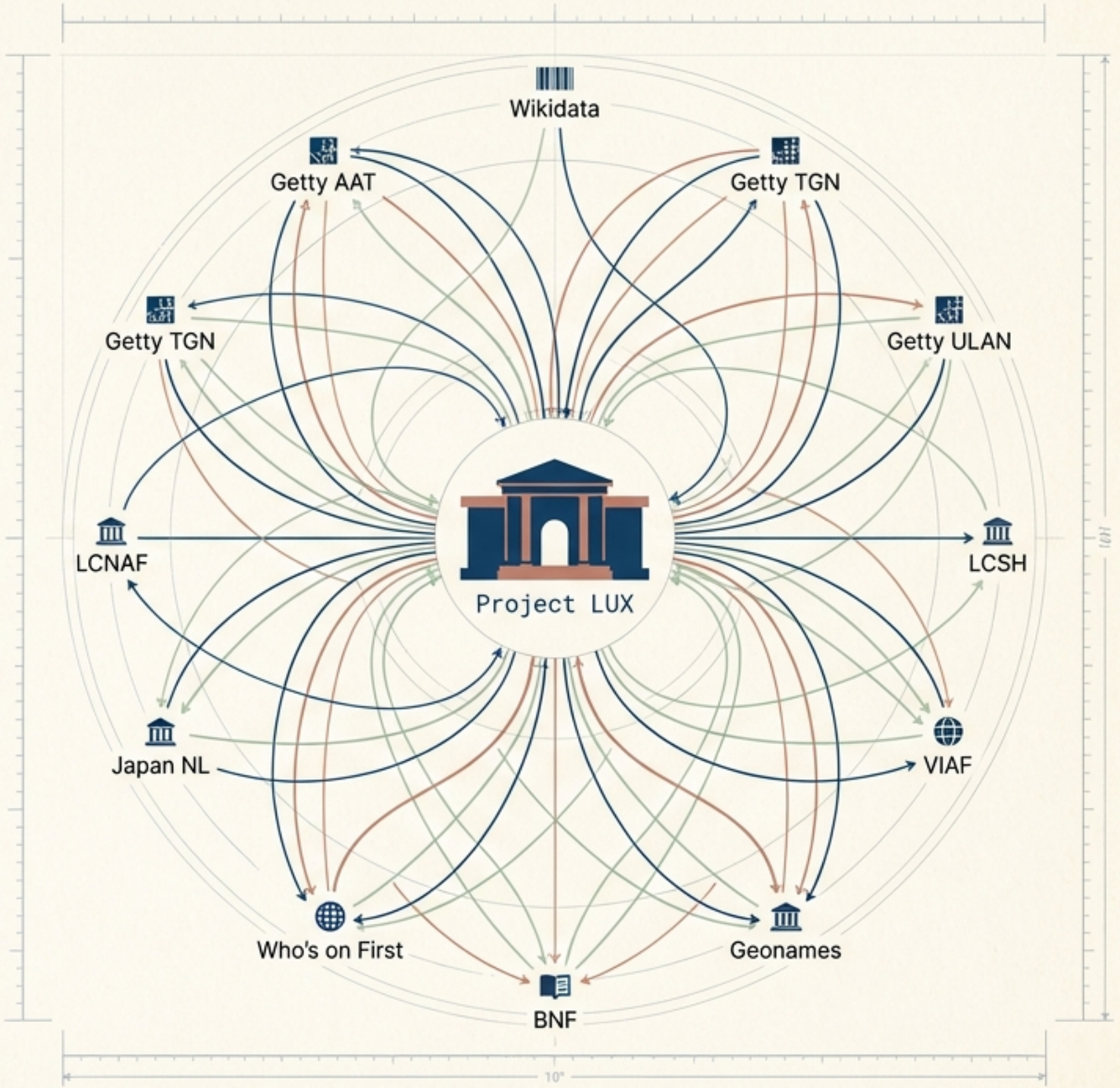
Once processed, the **Mapper** component translates the data into the standardized Linked Artformat. It can also export this enriched data directly into discovery layer formats—like MarkLogic or QLever—so downstream search tools and platforms can instantly understand and query the knowledge base.



NotebookLM

Integrating a global knowledge base

The pipeline pulls from a vast universe of external authorities to create its interconnected dataset. Integrations span across Wikidata, Getty (AAT, TGN, ULAN), LCNAF, LCSH, VIAF, Geonames, Who's on First, and numerous national libraries (Japan NL, BNF, BNE).

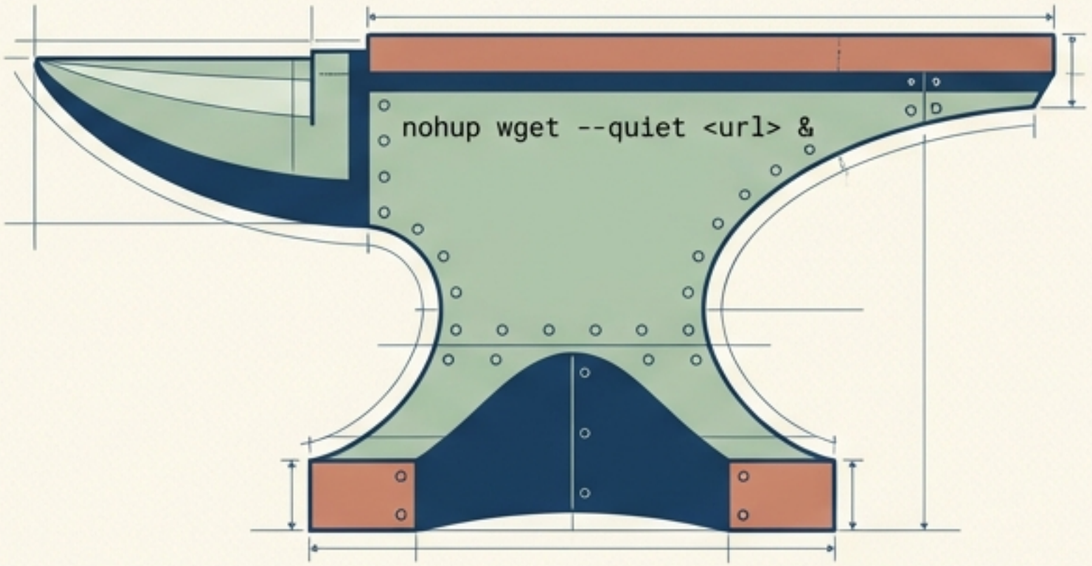


NotebookLM

Engineering pragmatism handles massive scale

The pipeline adapts its approach based on the source. For Getty's raw NTriples, it uses IIIF Change Discovery to efficiently sync only what's new. For massive datasets, pragmatism rules: Wikidata is fetched via background commands, and the staggering 33GB Who's on First SQLite file is utilized directly as a native geographic cache.

IIIF Change Discovery



33GB SQLite Cache

NotebookLM

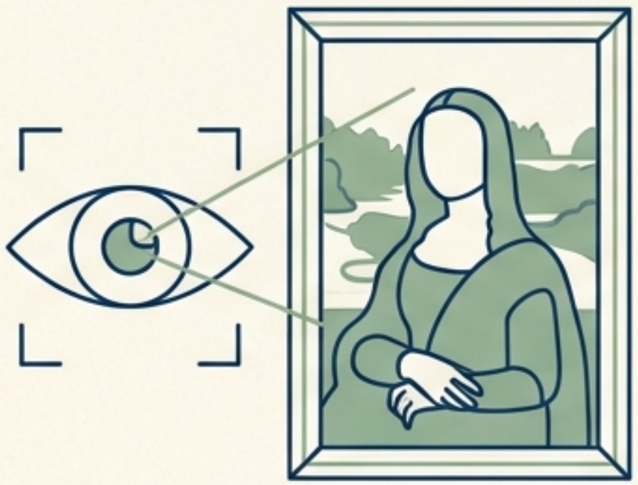
Shifting focus from wrangling to deep analysis

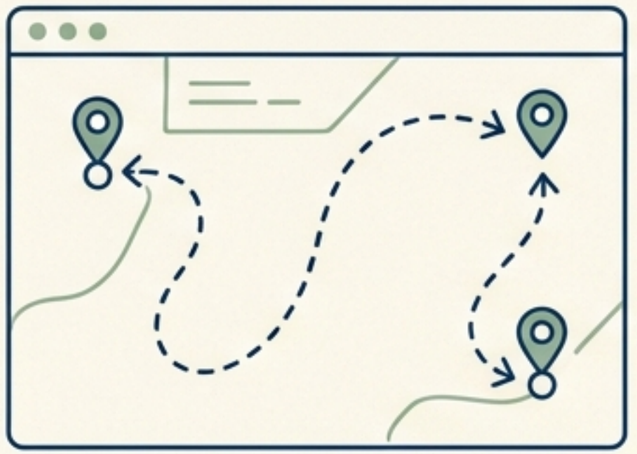
By automating the tedious cross-referencing, researchers are handed a distilled, interconnected knowledge base. They can abandon manual data cleanup and start asking complex, macro-historical questions—such as tracking how political events in 17th-century Amsterdam directly influenced the subject matter of art across Northern Europe.

NotebookLM

Clean data unlocks dynamic public experiences

A robust foundation of cleaned, linked data is the prerequisite for modern applications. This pipeline enables AI tools that can analyze artistic styles at scale, powers highly accurate virtual reality historical experiences, and allows the public to visually trace an artwork's journey across centuries and continents.

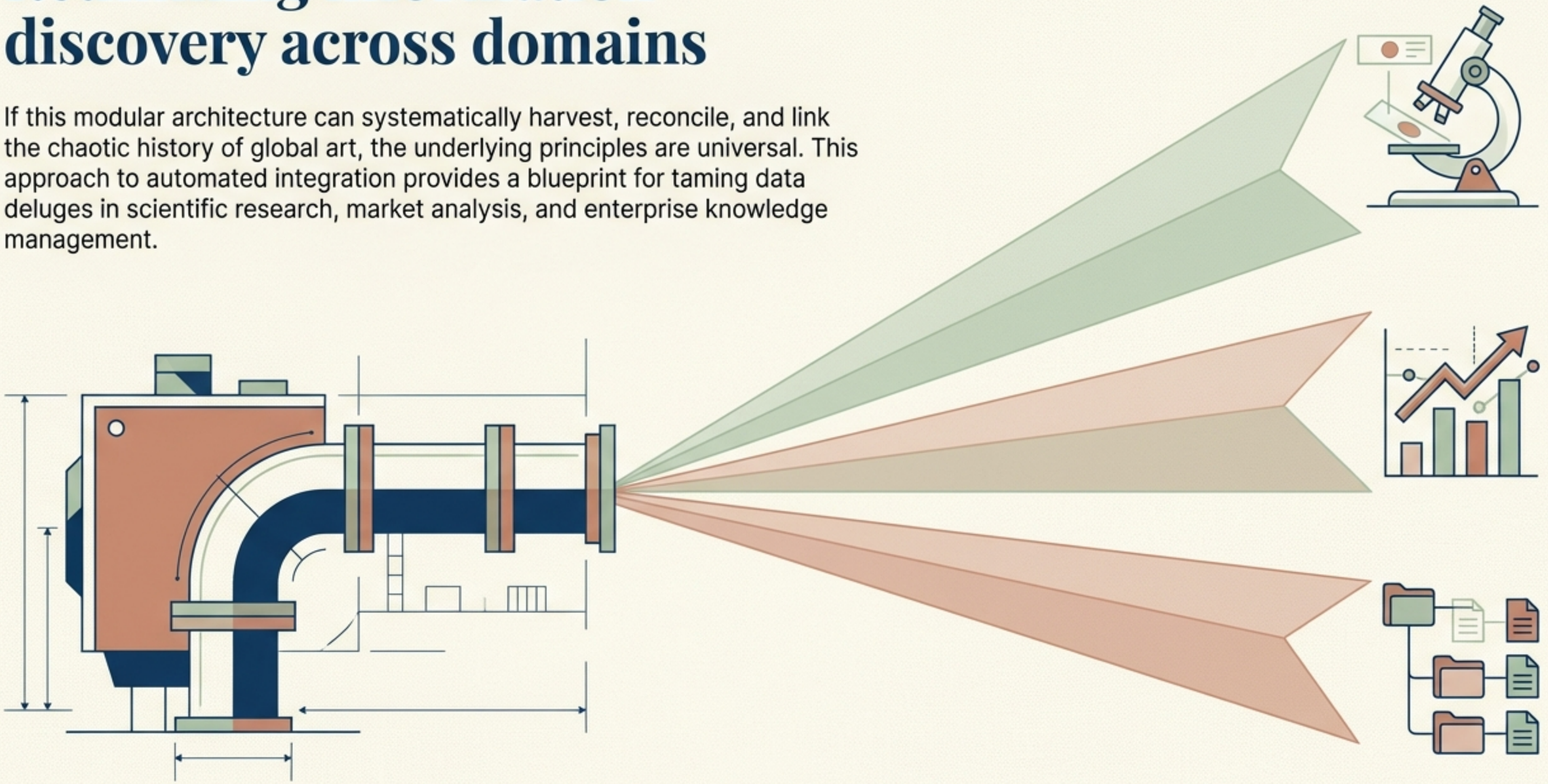




NotebookLM

Rethinking information discovery across domains

If this modular architecture can systematically harvest, reconcile, and link the chaotic history of global art, the underlying principles are universal. This approach to automated integration provides a blueprint for taming data deluges in scientific research, market analysis, and enterprise knowledge management.



NotebookLM